A Reliable Basis for Approximate Association Rules

Yue Xu, Yuefeng Li, Gavin Shaw

***Abstract*—For most of the work done in developing association rule mining, the primary focus has been on the efficiency of the approach and to a lesser extent the quality of the derived rules has been emphasized. Often for a dataset, a huge number of rules can be derived, but many of them can be redundant to other rules and thus are useless in practice. The extremely large number of rules makes it difficult for the end users to comprehend and therefore effectively use the discovered rules and thus significantly reduces the effectiveness of rule mining algorithms. If the extracted knowledge can't be effectively used in solving real world problems, the effort of extracting the knowledge is worth little. This is a serious problem but not yet solved satisfactorily. In this paper, we propose a concise representation called Reliable Approximate basis for representing non-redundant approximate association rules. We prove that the redundancy elimination based on the proposed basis does not reduce the belief to the extracted rules. We also prove that all approximate association rules can be deduced from the Reliable Approximate basis. Therefore the basis is a lossless representation of approximate association rules.**

***Index Terms*—Non-redundant association rule mining, approximate association rules, closed itemsets, certainty factor.**

I. Introduction

One big problem in association mining is the huge amount of the extracted rules which severely hinders the effective use of the discovered knowledge. Moreover, many of the extracted rules produce no value to the user or can be replaced by other rules thus considered redundant. Many efforts have been made on reducing the size of the extracted rule set. The approaches can be roughly divided to two categories, subjective approach and objective approach. In the subjective approach category, one technique is to define various interestingness measures and only the rules which are considered interesting based on the interesting measurements are generated[2], [3]. Another technique in this category is to apply constraints or templates to generate only those rules that satisfy the constraints or templates [1], [8], [11], [15]. In the objective approach category, the main technique is to construct concise representative bases of association rules without using user-dependent constraints. A concise representative basis contains much smaller number of rules and is considered lossless since all association rules can be derived from the basis. A number of concise representations of frequent patterns have been proposed, one of them, namely the closed itemsets, is of particular interest as they can be applied for generating non-redundant rules [9], [12], [19]. The notion of closed frequent itemset has its origins in the mathematical theory of Formal Concept Analysis introduced in the early 80s'[5], [16]. The use of frequent closed itemsets can greatly reduce the number of extracted rules and also provides a concise representation of association rules [13], [20]. Even though the number of extracted rules can be reduced by only using frequent closed itemsets, however, a considerable amount of redundancy still remains.

Yue Xu, Yuefeng Li and Gavin Shaw are with the Faculty of Information Technology, Queensland University of Technology, Brisbane, QLD4001, Australia (e-mail: {yue.xu,y2.li,gavin.shaw}@qut.edu.au).

Rules with confidence less than 1 are called Approximate rules and rules with confidence equal to 1 are called Exact rules. We have proposed a method to extract non-redundant exact rules [17]. In this paper, we present a concise representation basis called Reliable Approximate basis to extract non-redundant approximate rules. Most importantly, in this paper, we show that the redundancy elimination based on the proposed basis will not reduce the inference capacity of the extracted non-redundant rules. The certainty factor (CF) is an important and popular used measure of belief to inference rules [14]. We prove that the redundant rules eliminated by our approach have less or equal CF belief values than that of their corresponding non-redundant rules, and thus the elimination of such rules will not reduce the belief to the extracted rules. Moreover, we prove that all approximate association rules can be deduced from the reliable approximate basis, thus the reliable approximate basis is a lossless representation of approximate association rules.

The paper is organized as follows. Section II discusses redundancy in association rules and the elimination of the redundancy. Section III firstly introduces the proposed reliable basis for extracting non-redundant approximate rules, then presents a method to derive all approximate rules from the reliable basis. Experimental results are given in Section IV. Section V briefly discusses some related work. Finally, Section VI concludes the paper.

II. Redundancy and Redundancy Elimination

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct items, t be a transaction that contains a set of items such that $t \subseteq I$, T be a database containing different identifiable transactions. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called itemsets, and $X \bigcap Y = \emptyset$. Association rule mining is to find out association rules that satisfy the predefined minimum support (denoted as minsupp)and confidence (denoted as mincof) from a given database. The problem is usually decomposed into two subproblems: to find frequent itemsets and to generate association rules from those frequent itemsets. For the popular

TABLE I
ASSOCIATION RULES (MUSHROOM DATASET, MINSUPP=0.8, MINCONF=0.8)

	Rules (supp, conf)
1	gill-attachment-f ⇒ veil-type-p (0.97415,1.0)
2	veil-color-w ⇒ veil-type-p (0.97538 ,1.0)
3	gill-attachment-f,veil-color-w ⇒ veil-type-p (0.97317,1.0)
4	gill-attachment-f,ring-number-o ⇒ veil-type-p (0.89808,1.0)
5	gill-spacing-c,veil-color-w ⇒ veil-type-p (0.81487,1.0)
6	gill-attachment-f,gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81265,1.0)
7	gill-attachment-f,gill-spacing-c⇒ veil-type-p (0.81265,1.0)
8	gill-attachment-f,gill-spacing-c,veil-type-p ⇒ veil-color-w (0.81265 ,1.0)
9	gill-attachment-f ⇒ veil-type-p,veil-color-w (0.97317,0.99899)
10	gill-attachment-f ⇒ veil-type-p,ring-number-o (0.89808,0.92191)
11	veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81487,0.83544)
12	veil-color-w ⇒ gill-attachment-f,gill-spacing-c,veil-type-p (0.81265,0.83317)
13	gill-attachment-f,veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81265,0.83506)
14	gill-attachment-f,veil-color-w ⇒ veil-type-p,ring-number-o (0.8971,0.92183)
15	gill-attachment-f,ring-number-o ⇒ veil-type-p,veil-color-w (0.8971,0.9989)
16	gill-spacing-c,veil-color-w ⇒ gill-attachment-f,veil-type-p (0.81265,0.99728)
17	gill-attachment-f ⇒ veil-color-w (0.97317,0.99899)
18	gill-attachment-f ⇒ ring-number-o (0.89808,0.92191)
19	gill-attachment-f,veil-color-w ⇒ gill-spacing-c (0.81265,0.83506)
20	gill-attachment-f,ring-number-o ⇒ veil-color-w (0.8971,0.9989)

TABLE II
CLOSED ITEMSETS AND MINIMAL GENERATORS (MUSHROOM DATASET, MINSUPP=0.8)

Closed itemsets	Minimal Generators	Support
{ veil-type-p }		1.0
{gill-attachment-f,veil-type-p}	{gill-attachment-f}	0.97415
{gill-spacing-c,veil-type-p}	{gill-spacing-c}	0.8385
{veil-type-p,veil-color-w}	{veil-color-w}	0.97538
{veil-type-p,ring-number-o}	{ring-number-o}	0.9217
{gill-attachment-f, veil-type-p,veil-color-w}	{gill-attachment-f, veil-color-w}	0.97317
{gill-attachment-f,veil-type-p, ring-number-o}	{gill-attachment-f, ring-number-o}	0.8981
{gill-spacing-c,veil-type-p, veil-color-w}	{gill-spacing-c, veil-color-w}	0.81487
{gill-attachment-f,gill-spacing-c, veil-type-p,veil-color-w}	{gill-attachment-f, gill-spacing-c}	0.81265
{gill-attachment-f,veil-type-p, veil-color-w,ring-number-o}	{veil-color-w, ring-number-o}	0.8971

used Mushroom dataset (http://kdd.ics.uci.edu/), with minimal support 0.8 and minimal confidence 0.8, we can generate 88 association rules, 20 of them are displayed in Table I.

The definition of closed itemsets comes from the closure operation of the Galois connection [5]. Let I denote the set of items and T denote the set of transactions, 2^I and 2^T are the power set of I and T, respectively. $\forall i \in I$ and $\forall t \in T$, if item i appears in transaction t, then i and t has a binary relation δ denoted as $i\delta t$. The Galois connection of the binary relation is defined by the following mappings where $X \subseteq I$, $Y \subseteq T$:

$$\tau : 2^I \to 2^T, \tau(X) = \{t \in T | \forall i \in X, i\delta t\} \quad (1)$$

$$\gamma : 2^T \to 2^I, \gamma(Y) = \{i \in I | \forall t \in Y, i\delta t\} \quad (2)$$

$\tau(X)$ is called the transaction mapping of X. $\gamma(Y)$ is called the item mapping of Y. $\gamma \circ \tau(X)$, called the closure of X, gives the common items among the transactions each of which contains X.

Definition 1: (Closed Itemsets) Let X be a subset of I. X is a closed itemset iff $\gamma \circ \tau(X) = X$.

Definition 2: (Generators) An itemset $g \in 2^I$ is a generator of a closed itemset $c \in 2^I$ iff $c = \gamma \circ \tau(g)$ and $g \subset \gamma \circ \tau(g)$. g is said a minimal generator of the closed itemset set c if $\nexists g' \subset g$ such that $\gamma \circ \tau(g') = c$.

For the Mushroom dataset, the closed itemsets and their minimal generators (minsupp=0.8) are given in Table II.

A challenge to association mining is the huge amount of the extracted rules. Recent studies have shown that using closed itemsets and generators to extract association rules can greatly reduce the number of extracted rules [13], [19]. However, considerable amount of redundancy still exists in the extracted association rules extracted based on closed itemsets. In this section, firstly some examples are given to show the existence of redundancy in the extracted rules, then we define the redundancy to be removed, and at the end of this section we prove that the elimination of the defined redundancy won't reduce the belief to the extracted non-redundant rules. In Section 3, we describe a concise representation of the defined non-redundant association rules, from which all approximate association rules can be derived.

A. Redundancy Definition

The rules in Table I are considered useful based on the predefined minimum support and confidence. However, some of the rules actually do not contribute new information. The consequent concluded by some rules can be obtained from some other rules without requiring more conditions but with higher or the same confidences. For example, in order to be fired the rules 5, 8, 13, and 20 in Table I require more conditions than that of rules 2, 6, 11, and 9, respectively, but conclude the same or less results which can be produced by rules 2, 6, 11, and 9. That means, without rules 5, 8, 13, and 20, we still can achieve the same result using other rules. Therefore, rules 5, 8, 13, and 20 are considered redundant to rules 2, 6, 11, and 9, respectively. Comparing to rules 2, 6, 11, and 9, the redundant rules 5, 8, 13, and 20 have a longer or the same antecedent and a shorter or the same consequent, respectively, and the confidence of the redundant rules is not larger than that of their corresponding non-redundant rules. The following definition defines such kind of redundant rules.

Definition 3: (Redundant rules) Let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ be two association rules with confidence cf and cf', respectively. $X \Rightarrow Y$ is said a redundant rule to $X' \Rightarrow Y'$ if $X' \subseteq X$, $Y \subseteq Y'$, and $cf \leq cf'$.

Based on Definition 3, for an association rule $X \Rightarrow Y$, if there does not exist any other rule $X' \Rightarrow Y'$ such that the confidence of $X' \Rightarrow Y'$ is the same as or larger than the confidence of $X \Rightarrow Y$, $X' \subseteq X$ or $Y \subseteq Y'$, then $X \Rightarrow Y$ is non-redundant. Definition 3 is similar to Pasquier's definition of min-max association rules [13]. However, Pasquier's definition requires that a redundant rule and its corresponding non-redundant rule must have identical confidence and identical support, while Definition 3 here only requires that the confidence of the redundant rule is not larger than that of its corresponding non-redundant rule. In the following subsection, we prove that the requirement relaxation to redundancy will not reduce the belief to the extracted non-redundant rules.

B. Redundancy Elimination

The certainty factor theory were first introduced in MYCIN [14] to express how accurate and truthful a rule is and how reliable the antecedent of the rule is. The certainty factor theory is based on two functions: measure of belief $MB(X,Y)$ and measure of disbelief $MD(X,Y)$ for a rule $X \Rightarrow Y$, as given below.

$$MB(X,Y) = \begin{cases} 1 & P(Y) = 1 \\ 0 & P(Y/X) \leq P(Y) \\ \frac{P(Y/X)-P(Y)}{1-P(Y)} & otherwise \end{cases} \tag{3}$$

$$MD(X,Y) = \begin{cases} 1 & P(Y) = 0 \\ 0 & P(Y/X) \geq P(Y) \\ \frac{P(Y)-P(Y/X)}{P(Y)} & otherwise \end{cases} \tag{4}$$

where, in the context of association rules, $P(Y/X)$ and $P(Y)$ are the confidence of the rule and the support of the consequent, respectively. The values of $MB(X,Y)$ and $MD(X,Y)$ range between 0 and 1 measuring the strength of belief or disbelief in consequent Y given antecedent X. $MB(X,Y)$ weighs how much the antecedent X increases the possibility of Y occurring. If the antecedent completely support the consequent, then $P(Y/X)$ will equal to 1 thus $MB(X,Y)$ will be 1. $MD(X,Y)$=1 indicates that the antecedent completely denies the consequent thus the disbelief in the rule reaches its highest value. The total strength of belief or disbelief in the association captured by the rule is measured by the certainty factor which is defined as follows:

$$CF(X,Y) = MB(X,Y) - MD(X,Y) \tag{5}$$

The value of a certainty factor is between 1 and -1. Negative values represent cases where the antecedent is against the consequent; positive values represent that the antecedent supports the consequent; while CF=0 means that the antecedent does not influence the belief to Y. Obviously, association rules with high CF values are more useful since they represent strong positive associations between antecedents and consequents. Indeed, the aim of association rule mining is to discover strong positive associations from large amount of data. Therefore, we propose that the certainty factors can be used to measure the strength of discovered association rules.

Theorem 1 below states that the CF value of a redundant rule defined by Definition 3 will never be larger than the CF value of its corresponding non-redundant rules. It means that, the association between the antecedent and consequent of the non-redundant rule is stronger than that of the redundant rule.

Theorem 1: Let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ be two association rules. If $Y' \subseteq Y$, and $P(Y/X) \geq P(Y'/X')$, then $CF(X,Y) \geq CF(X',Y')$.

Proof: From Equation (5) we have

$CF(X,Y)$ - $CF(X',Y')$

$= MB(X,Y)$-$MB(X',Y')$+$MD(X',Y')$-$MD(X,Y)$

1) Assuming that $P(Y'/X') \geq P(Y')$. From condition $Y' \subseteq Y$, we have $P(Y) \leq P(Y')$. Because $P(Y/X) \geq P(Y'/X')$, so we have $P(Y/X) \geq P(Y)$. In this case, $MD(X',Y')$-$MD(X,Y)$ = 0. To prove the theorem, we need to prove that $MB(X,Y)$ - $MB(X',Y') \geq 0$. From Equation (3), we have:

$MB(X,Y)$ - $MB(X',Y')$ $=\frac{P(Y/X)-P(Y)}{1-P(Y)} - \frac{P(Y'/X')-P(Y')}{1-P(Y')}$

$=\frac{(P(Y/X)-P(Y))(1-P(Y'))-(P(Y'/X')-P(Y'))(1-P(Y))}{(1-P(Y))(1-P(Y'))}$

$=\frac{P(Y/X)-P(Y'/X')+P(Y'/X')P(Y)-P(Y/X)P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$

Let $\alpha = P(Y/X) - P(Y'/X')$, the above expression becomes:

$=\frac{\alpha+P(Y'/X')P(Y)-(\alpha+P(Y'/X'))P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$

$=\frac{\alpha+P(Y'/X')P(Y)-\alpha P(Y')-P(Y'/X')P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$

$=\frac{\alpha(1-P(Y'))+P(Y'/X')(P(Y)-P(Y'))-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$

$=\frac{\alpha(1-P(Y'))+(P(Y')-P(Y))(1-P(Y'/X'))}{(1-P(Y))(1-P(Y'))}$

Because $P(Y) \leq P(Y')$ and $P(Y/X) \geq P(Y'/X')$ which makes $\alpha \geq 0$, we prove that the above expression ≥ 0. Hence, $MB(X,Y)$ - $MB(X',Y') \geq 0$

2) Assuming that $P(Y'/X') \leq P(Y')$. In this situation, we have two cases.

(i) $P(Y/X) \leq P(Y)$

In this case, $MB(X,Y)$-$MB(X',Y')$ = 0. To prove the theorem, we need to prove that $MD(X',Y')$-$MD(X,Y) \geq 0$. From Equation (4), we have

$MD(X',Y')$-$MD(X,Y)$ $=\frac{P(Y')-P(Y'/X')}{P(Y')} - \frac{P(Y)-P(Y/X)}{P(Y)}$

After expanding the above expression and eliminating identical dual terms, we have

$MD(X',Y')$-$MD(X,Y)$ $=\frac{P(Y/X)P(Y')-P(Y'/X')P(Y)}{P(Y)P(Y')}$.

$\geq \frac{P(Y/X)P(Y')-P(Y/X)P(Y)}{P(Y)P(Y')}$.

Again, since $P(Y) \leq P(Y')$, we get

$MD(X',Y')$-$MD(X,Y) \geq 0$.

(ii) $P(Y/X) \geq P(Y)$

In this case, $MD(X,Y)$=0 and $MB(X',Y')$ = 0. To prove the theorem, we need to prove that $MD(X',Y')$+$MB(X,Y) \geq 0$. Because $P(Y'/X') \leq P(Y')$ and $P(Y/X) \geq P(Y)$, from the equations (3) and (4), it is true that

$MD(X',Y')$+$MB(X,Y) \geq 0$

Combining the results of the above cases, we have

$CF(X,Y)$ - $CF(X',Y') \geq 0$, hence

$CF(X,Y) \geq CF(X',Y')$

■

According to Theorem 1, the CF value of a redundant rule defined by Definition 3 is never higher than that of its corresponding non-redundant rule and thus the elimination of such redundant rules is reliable since it won't reduce the belief to the extracted non-redundant rules.

III. Concise Basis for Non-redundant Approximate Association Rules

Pasquier et al. [13] proposed a condensed basis to represent non-redundant approximate association rules, which is defined as follows:

Definition 4: (Min-max Approximate Basis) Let C be the set of frequent closed itemsets and G be the set of minimal generators of the frequent closed itemsets in C. The min-max approximate basis is:

$$MinMaxApprox = \{g \Rightarrow (c \backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c\}$$

For the 88 rules extracted from the Mushroom dataset mentioned above, there are 71 approximate rules. Based on the Min-max approximate basis, 25 approximate rules, as displayed in Table III, are extracted and considered non-redundant in terms of the redundancy definition given in [13]. However, under Definition 3, some of the rules extracted from the min-max approximate basis are redundant such as rules 22 to 25 which are redundant to rules 17, 11, 10, and 16, respectively.

TABLE III
NON-REDUNDANT APPROXIMATE RULES EXTRACTED FROM MIN-MAX APPROXIMATE BASIS (MUSHROOM DATASET, MINSUPP=0.8, MINCONF=0.8)

	Rules (supp, conf)
1	veil-type-p $\Rightarrow$ gill-attachment-f (0.97415,0.97415)
2	veil-type-p $\Rightarrow$ gill-spacing-c (0.8385 ,0.8385)
3	veil-type-p $\Rightarrow$ veil-color-w (0.97538,0.97538)
4	veil-type-p $\Rightarrow$ ring-number-o (0.92171,0.92171)
5	veil-type-p $\Rightarrow$ gill-attachment-f,veil-color-w (0.97317,0.97317)
6	veil-type-p $\Rightarrow$ gill-attachment-f, ring-number-o (0.89808,0.89808)
7	veil-type-p $\Rightarrow$ gill-spacing-c, veil-color-w (0.81487,0.81487)
8	veil-type-p $\Rightarrow$ gill-attachment-f,gill-spacing-c, veil-color-w (0.81265,0.81265)
9	veil-type-p $\Rightarrow$ gill-attachment-f,veil-color-w, ring-number-o (0.8971,0.8971)
10	gill-attachment-f $\Rightarrow$ veil-type-p, veil-color-w (0.97317, 0.99899)
11	gill-attachment-f $\Rightarrow$ veil-type-p, ring-number-o (0.89808,0.92191)
12	gill-attachment-f $\Rightarrow$ gill-spacing-c,veil-type-p, veil-color-w (0.81265,0.83422)
13	gill-attachment-f $\Rightarrow$ veil-type-p,veil-color-w, ring-number-o (0.8971,0.9209)
14	gill-spacing-c $\Rightarrow$ veil-type-p,veil-color-w (0.81487,0.97181)
15	gill-spacing-c $\Rightarrow$ gill-attachment-f,veil-type-p, veil-color-w (0.81265,0.96917)
16	veil-color-w $\Rightarrow$ gill-attachment-f, veil-type-p (0.97317, 0.99773)
17	veil-color-w $\Rightarrow$ gill-spacing-c,veil-type-p (0.81487,0.83544)
18	veil-color-w $\Rightarrow$ gill-attachment-f,gill-spacing-c, veil-type-p(0.81265, 0.83317)
19	veil-color-w $\Rightarrow$ gill-attachment-f,veil-type-p, ring-number-o (0.8971,0.91974)
20	ring-number-o $\Rightarrow$ gill-attachment-f, veil-type-p (0.89808, 0.97436)
21	ring-number-o $\Rightarrow$ gill-attachment-f,veil-type-p, veil-color-w (0.8971, 0.97329)
22	gill-attachment-f,veil-color-w $\Rightarrow$ gill-spacing-c, veil-type-p (0.81265,0.83506)
23	gill-attachment-f,veil-color-w $\Rightarrow$ veil-type-p, ring-number-o (0.8971, 0.92183)
24	gill-attachment-f,ring-number-o $\Rightarrow$ veil-type-p, veil-color-w (0.8971,0.9989)
25	gill-spacing-c,veil-color-w $\Rightarrow$ gill-attachment-f, veil-type-p (0.81265, 0.99728)

A. Reliable Approximate Basis

Corresponding to the Min-max approximate basis, we propose a more concise basis called Reliable Approximate basis as defined in Definition 5.

Definition 5: (Reliable Approximate Basis) Let C be the set of frequent closed itemsets and G be the set of minimal generators of the frequent closed itemsets in C. The Reliable approximate basis is:

$$\begin{aligned} &ReliableApprox \\ &= \{g \Rightarrow (c\backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c, \neg(g \supseteq ((c\backslash c') \cup g')) \\ &\qquad \text{or } conf(g \Rightarrow (c\backslash g)) > conf(g' \Rightarrow (c'\backslash g')) \\ &\qquad where\ c' \in C, g' \in G, g' \subset g, \gamma \circ \tau(g') \subset c'\} \end{aligned}$$

The correctness of the above definition can be proved by the following theorems and properties.

Lemma 1: Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma\circ\tau(g) \subset c$. If $\exists c' \in C$, $\exists g' \in G$, $\gamma\circ\tau(g') \subset c'$, $g' \subset g$, $g \supseteq ((c\backslash c')\cup g')$, and $conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g')$, then $g \Rightarrow c\backslash g$ is redundant to $g' \Rightarrow c'\backslash g'$.

Proof: Let $A = c\backslash c'$ so that $c \subseteq A \cup c'$ and $A \cap c' = \emptyset$. Therefore, we have $c\backslash((c\backslash c') \cup g') \subseteq (A \cup c')\backslash(A \cup g')$. From $\gamma \circ \tau(g') \subset c'$, we have $g' \subset c'$. Since $A \cap c' = \emptyset$, then $A \cap g' = \emptyset$. So, $c\backslash((c\backslash c')\cup g') \subseteq (A\cup c')\backslash(A\cup g') = ((A\cup c')\backslash A)\backslash g') = c'\backslash g'$. That is, $c\backslash((c\backslash c') \cup g') \subseteq c'\backslash g'$. Because $g \supseteq ((c\backslash c') \cup g')$, we have $c\backslash g \subseteq c\backslash((c\backslash c') \cup g') \subseteq c'\backslash g'$, hence, $c\backslash g \subseteq c'\backslash g'$. Since $c\backslash g \subseteq c'\backslash g'$, $g \supset g'$, and $conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g')$, according to Definition 3, we can conclude that $g \Rightarrow c\backslash g$ is redundant to $g' \Rightarrow c'\backslash g'$. ∎

According to Modus tolen inference rule, i.e., if the consequent of an implication is false, the antecedent of the rule must be false, from Lemma 1, we get the following corollary:

Corollary 1: Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma \circ \tau(g) \subset c$. If $g \Rightarrow c\backslash g$ is a non-redundant rule, then $\forall c' \in C$, $\forall g' \in G$, $\gamma\circ\tau(g') \subset c'$ and $g' \subset g$, we have $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$.

Theorem 2: Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma\circ\tau(g) \subset c$. $g \Rightarrow c\backslash g$ is a non-redundant rule iff $\forall c' \in C$, $\forall g' \in G$, $g' \subset g$, $\gamma\circ\tau(g') \subset c'$, and $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$.

Proof:

1) $\Longrightarrow$. The proof follows the conclusion of Corollary 1.
2) $\Longleftarrow$. (i) Assuming that $\neg(g \supseteq ((c\backslash c') \cup g'))$, we get $g \subset (c\backslash c') \cup g'$, or $g \cap ((c\backslash c') \cup g') = \emptyset$, or $(g\cap((c\backslash c')\cup g') \subset ((c\backslash c')\cup g'))\wedge(g\cap((c\backslash c')\cup g') \subset g)$.

(1). In the case that $g \subset (c\backslash c') \cup g'$ is true, assuming that $g \Rightarrow c\backslash g$ is redundant, then we get, $\exists c' \in C$, $\exists g' \in G$, and $\gamma \circ \tau(g') \subset c'$ (hence $g' \subset c'$) such that $g' \subseteq g$ and $c'\backslash g' \supseteq c\backslash g$.
From $c'\backslash g' \supseteq c\backslash g$ and $g' \subseteq c'$, we have $c' \supseteq c'\backslash g' \supseteq c\backslash g$, i.e., $c' \supseteq c\backslash g$. Since $\gamma \circ \tau(g) \subset c$ thus $g \subset c$, obviously we have $c = (c\backslash g) \cup g$ and $(c\backslash g)\cap g = \emptyset$; also $(c\backslash c')\cup c' \supseteq c$ and $(c\backslash c')\cap c' = \emptyset$ are true. Therefore, we have $(c\backslash c')\cup c' \supseteq c = (c\backslash g)\cup g$, i.e.:

$$(c\backslash c') \cup c' \supseteq (c\backslash g) \cup g \qquad (a)$$

Because $c' \supseteq c\backslash g$, $(c\backslash c') \cap c' = \emptyset$ and $(c\backslash g) \cap g = \emptyset$, after c' being removed from the left side of (a) and $c\backslash g$ being removed from the right side of (a), the formula (a) becomes $c\backslash c' \subseteq g$. From $g' \subseteq g$, we get $(c\backslash c') \cup g' \subseteq g \cup g' = g$, i.e., $(c\backslash c') \cup g' \subseteq g$ which contradicts to $(c\backslash c') \cup g' \supset g$.
Therefore, the assumption is false, i.e., $g \Rightarrow c\backslash g$ is non-redundant.

(2). In the case that $g \cap ((c\backslash c') \cup g') = \emptyset$ is true, $g \cap g' = \emptyset$, thus $g \supset g'$ is always false. Therefore, $g \Rightarrow c\backslash g$ can't be redundant to $g' \Rightarrow c'\backslash g'$.

(3). In the case that $(g \cap ((c\backslash c') \cup g') \subset ((c\backslash c') \cup g')) \wedge (g \cap ((c\backslash c') \cup g') \subset g)$ is true, there must exist some x such that $x \in c\backslash c'$ and $x \notin g$ or $x \in g'$ and $x \notin g$. The former will make $(c\backslash g) \subset (c'\backslash g')$ false and the latter will make $g \supset g'$ false. Therefore, $g \Rightarrow c\backslash g$ will never be redundant to $g' \Rightarrow c'\backslash g'$

(ii) Assuming that $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$. From Definition 3, we can directly conclude that $g \Rightarrow c\backslash g$ is not redundant.

■

The proposed Reliable Approximate Basis defines a more concise set of approximate rules which are non-redundant, sound and lossless. The algorithm to extract non-redundant approximate rules based on the Reliable Approximate Basis is given below:

Algorithm 1: **ReliableApproxBasis**($Closure$)

Input: $Closure$: a set of frequent closed itemsets
$Generator$: a set of minimal generators

Output: A set of non-redundant approximate rules.

1. $approxRules := \emptyset$
2. for each $c \in Closure$
3. for each $g \in Generator$ such that $\gamma \circ \tau(g) \subset c$
4. if $\forall c' \in Closure$, $\forall g' \in G$ such that $\gamma \circ \tau(g') \subset c'$ and $g' \subseteq g$
5. we have $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$
6. then $approxRules := approxRules \cup \{g \Rightarrow (c\backslash g)\}$
7. end-for
8. end-for
9. Return $appproxRules$

For the Mushroom example dataset, 21 non-redundant approximate rules are extracted based on the Reliable Approximate basis. Rules 22, 23, 24 and 25 in Table III extracted based on the Minmax Approximate basis are considered redundant under the Reliable Approximate basis, respectively, and thus eliminated.

B. Deriving All Approximate Association Rules

Algorithms have been proposed to derive all association rules from the Min-max bases [13] and the Reliable Exact basis [17]. In this section, we provide an algorithm that can derive all approximate rules from the Reliable Approximate basis.

According to the definitions 4 and 5, the Min-max Approximate basis can be described as:

$$
\begin{aligned}
&MinMaxApprox = \{g \Rightarrow (c\backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c\}\\
&= \{g \Rightarrow (c\backslash g) | c \in C, g \in G,\\
&\quad (\neg(g \supseteq ((c\backslash c') \cup g')) \text{ or } conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g'))\\
&\quad for\ all\ c' \in C, g' \in G, \gamma \circ \tau(g) \subset c\\
&\quad or\ \ (g \supseteq ((c\backslash c') \cup g') \text{ and } conf(g \Rightarrow c\backslash g) \le conf(g' \Rightarrow c'\backslash g'))\\
&\quad for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\}\\
&= \{g \Rightarrow (c\backslash g) | c \in C, g \in G_c,\\
&\quad \neg(g \supseteq ((c\backslash c') \cup g')) \text{ or } conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')\\
&\quad for\ all\ c' \in C, g' \in G, \gamma \circ \tau(g) \subset c\} \cup\\
&\quad \{g \Rightarrow (c\backslash g) | c \in C, g \in G_c,\\
&\quad g \supseteq ((c\backslash c') \cup g') \text{ and } conf(g \Rightarrow c\backslash g) \le conf(g' \Rightarrow c'\backslash g')\\
&\quad for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\}\\
&= ReliableApprox \cup NonReliableApprox
\end{aligned}
$$

Where

$$
\begin{aligned}
&NonReliableApprox = \{g \Rightarrow (c\backslash g) | c \in C, g \in G_c,\\
&g \supseteq ((c\backslash c') \cup g') \text{ and } conf(g \Rightarrow c\backslash g) \le conf(g' \Rightarrow c'\backslash g')\\
&for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\}
\end{aligned} \quad (6)
$$

The following theorem showes that, for $r_2 : g_2 \Rightarrow c_2\backslash g_2$, $c_2 \in C$ and $g_2 \in G$ (i.e., r_2 is a rule in $MinMaxApprox$), if for some $c_1 \in C$ and some $g_1 \in G$, there is $(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$ and $conf(r_1) \le conf(r_2)$, then we can deduce: $r_1 : g_1 \Rightarrow c_1\backslash g_1$ is a rule in $NonReliableApprox$. This means that, from a rule in $MinMaxApprox$, we could deduce a $NonReliableApprox$ rule.

Theorem 3: Let C be the set of frequent closed itemsets and G be the set of minimal generators. For rules $r_1 : g_1 \Rightarrow c_1\backslash g_1$ and $r_2 : g_2 \Rightarrow c_2\backslash g_2$ where c_1, $c_2 \in C$, g_1 , $g_2 \in G$, $\gamma \circ \tau(g_1) \subset c_1$, and $\gamma \circ \tau(g_2) \subset c_2$. r_1 is a NonReliable approximate rule iff $(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$ and $conf(r_1) \le conf(r_2)$.

Proof:

1) $\Longrightarrow$
According to the definition of Min-max approximate basis, both $r_1 : g_1 \Rightarrow c_1\backslash g_1$ and $r_2 : g_2 \Rightarrow c_2\backslash g_2$ are Min-max approximate rules. If $g_1 \supseteq (c_1\backslash c_2) \cup g_2$ and $conf(r_1) \le conf(r_2)$, then $\neg(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$ must be false. According to the definition of Reliable approx basis, $r_1 \notin ReliableApprox$ must be true. Therefore, $r_1 \in NonReliableApprox$ is true.

2) $\Longleftarrow$
Assuming that $r_1 : g_1 \Rightarrow c_1\backslash g_1 \in NonReliableApprox$. From Equation (6), we immediately get, $g_1 \supseteq ((c_1\backslash c_2) \cup g_2)$ and $conf(r_1) \le conf(r_2)$ for some $c_2 \in C$,, and $g2 \in G$.

■

We designed the following algorithm $AllApproxFromReliable$ to derive all approximate rules from the Reliable Approx basis. The algorithm $AllApproxFromReliable$ takes $ReliableApprox$ as the initial value for $MinMaxApprox$. Steps 4-8 generate approximate rules from an approximate basis rule in current $MinMaxApprox$. Steps 9 to 14 deduce $NonReliableApprox$ basis rules and add them into the current $MinMaxApprox$. Therefore, during the process of deriving approximate rules, we generates all $NonReliableApprox$ rules so that $MinMaxApprox$ will be completed progressively during the course. Theorem 3 ensures that we can deduce all $NonReliableApprox$ basis rules. On completion of executing Algorithm 2, $MinMaxApprox$ will contains all $ReliableApprox$ basis rules and also all $NonReliableApprox$ basis rules. Steps 17 to 21 in Algorithm 2 derive all approximate rules from these basis rules, which performs the same task as the steps 11 to 17 in the approximate reconstruction algorithm proposed in [13].

Algorithm 2: **AllApproxFromReliable**($ReliableApprox$)

Input: $ReliableApprox$: reliable approximate basis

Output: $AllApprox$: A set of all approximate association rules

1. $AllExact := \emptyset$, $MinMaxApprox := ReliableApprox$
2. for $i = 2$ to maximum size of closed itemsets
3. for rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp, r_1.conf) \in$

$MinMaxApprox$
and $|c_1| = i$
for subset $c_2 \subset c_1$
if $(r_2 : a_1 \Rightarrow c_2, r_2.supp, r_2.conf) \notin AllApprox$
and $r_2.conf \neq 1$//r_2 is not an exact rule
then $AllApprox := AllApprox \cup$
$\{(r_2 : a_1 \Rightarrow c_2, r_1.supp, r_1.conf)\}$
end-for
for each closed itemset c_3
for generator a such that $a \supseteq a_1$ and $a.closure \subset c_3$
if $a \supseteq ((c_3 \backslash (c_1 \cup a_1)) \cup a_1)$ and $r_1.conf \geq \frac{c_3.supp}{a.supp}$
then $MinMaxApprox := MinMaxApprox \cup$
$\{a \Rightarrow (c_3 \backslash a), c_3.supp, \frac{c_3.supp}{a.supp}\}$
end-for
end-for
end-for
end-for
for rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp, r_1.conf) \in AllApprox$
for each subset $c_3 \subseteq c_2$ where $c_2 = a_1.closure \backslash a_1$, $\frac{(a_1.closure).supp}{a_1.supp} = 1$
$AllApprox := AllApprox \cup$
$\{a_1 \cup c_3 \Rightarrow c_1 \backslash c_3, r_1.supp, r_1.conf\}$
end-for
end-for
return $AllExact$

IV. Experiments

We have conducted experiments to evaluate the effectiveness of the proposed Reliable approximate basis. This section presents the experimental results.

A. Datasets

We used the following three datasets from UCI KDD Archive (http://kdd.ics.uci.edu/). The Mushrooms dataset contains 8,124 records each of which describes the characteristics of one mushroom object. Each mushroom object has 23 attributes some of which are multiple value attributes. After converting the multiple value attributes to binary ones, the number of attributes of each object becomes 126. The Annealing dataset contains 898 annealing instances (objects), each has 38 attributes. After converting multiple value attributes to binary ones, each object has 276 attributes. The Flare2 dataset contains 1,066 solar flare instances each of which represents captured features for one active region on the sun. Each flare instance has 50 attributes after the multiple value attributes are converted to binary attributes. The experiment is to find the associations among attributes for the three datasets.

B. Evaluation Results

In this experiment, firstly we confirm that both the MinMax basis and the Reliable basis can deduce all approximate rules. For example, when $Minsupp$ is 0.3, both bases produce 21,377 approximate rules for the Mushroom dataset as showed in Table IV. Secondly, we test the reduction ratio between the size of the $MinMaxApprox$ basis and the size of the $ReliableApprox$ basis for different $Minsupp$ settings.

TABLE IV
Number of approximate rules (Mushroom dataset, minconf=0.5)

Minsupp	Approx rules derived (MinMax,Reliable)	MinMax Approx Basis	Reliable Approx Basis	Reduction Ratio
0.3	21,377	2,634	1,970	25%
0.4	2,528	465	361	22%
0.5	835	175	135	23%
0.6	228	59	52	12%
0.7	161	39	34	13%
0.8	71	25	21	16%

For all tests, the $minconf$ was set to 0.5. Table IV, Table V, and Table VI present the test results for the three datasets, respectively.

The experiment results showed that the reduction is considerable high. For instance, when $Minsupp$ was set to 0.3, for the Annealing dataset, the $MinMax$ basis contains 865 basis rules as showed in Table V, while the $Reliable$ basis contains 554 basis rules, the reduction ratio is 36%. In this case, 5,052 approximate rules can be deduced either from the $MinMax$ basis or from the $Reliable$ basis. For example, the following 9 rules in the $MinMax$ basis are redundant to the reliable rule *steel-A* $\Rightarrow$ *product-type-C,strength-000* (0.4844, 0.9886), therefore they are excluded in the $Reliable$ basis:

steel-A,carbon-00 $\Rightarrow$ *product-type-C,strength-000, (0.47327, 0.9884)*
steel-A,hardness-00 $\Rightarrow$ *product-type-C,strength-000, (0.30512,0.9821)*
steel-A,bore-0000 $\Rightarrow$ *product-type-C,strength-000, (0.4655,0.9882)*
steel-A,class-3 $\Rightarrow$ *product-type-C,strength-000, (0.3853,0.9858)*
steel-A,carbon-00,bore-0000 $\Rightarrow$ *product-type-C,strength-000, (0.4543, 0.9879)*
steel-A,carbon-00,class-3 $\Rightarrow$ *product-type-C,strength-000, (0.3775,0.9854)*
steel-A,hardness-00,bore-0000 $\Rightarrow$ *product-type-C,strength-000, (0.3040, 0.9820)*
steel-A,bore-0000,class-3 $\Rightarrow$ *product-type-C,strength-000, (0.3731,0.9853)*
steel-A,carbon-00,bore-0000,class-3 $\Rightarrow$ *product-type-C,strength-000, (0.3653,0.9850)*

The 9 rules listed above have the same consequent but a larger antecedent than that of the reliable rule *steel-A* $\Rightarrow$ *product-type-C,strength-000*. Both the support and confidence values , as indicated as (support, confidence) at the end of each rule, of these 9 rules are smaller than that of the reliable rule. Therefore, according to Theory 1, their CF value won't be greater than that of the reliable rule. In real world problem solving, if we know that *steel-A* is true, by applying the rule *steel-A* $\Rightarrow$ *product-type-C,strength-000*, we can conclude that *product-type-C,strength-000* is true. We don't have to know *hardness-00*, *class-3*, or *bore-0000*, etc. in order to reach this consequence. That means, all the 9 rules are useless or redundant if we have the rule *steel-A* $\Rightarrow$ *product-type-C,strength-000* at hand. Eliminating these redundant rules can greatly reduce the size of the discovered rule set, but the capacity of the rule base in solving problems remains the same.

V. Related Work

Many approaches have been proposed aiming at reducing the number of extracted rules and improving the"usefulness" of the rules as well[1], [3], [7], [15]. Also some work has been done on concisely representing and interpreting multidimensional association rules using granules and multi-tier

TABLE V
NUMBER OF APPROXIMATE RULES (ANNEALING DATASET, MINCONF=0.5)

Minsup	Approx rules derived (MinMax,Reliable)	MinMax Approx Basis	Reliable Approx Basis	Reduction Ratio
0.3	5,052	865	554	36%
0.4	1,835	435	296	32%
0.5	1,186	300	218	27%
0.6	416	137	102	26%

TABLE VI
NUMBER OF APPROXIMATE RULES (FLARE2 DATASET, MINCONF=0.5)

Minsupp	Approx rules derived (MinMax,Reliable)	MinMax Approx Basis	Reliable Approx Basis	Reduction Ratio
0.3	7,604	1216	710	42%
0.4	2,420	644	479	27%
0.5	5,599	1081	730	32%
0.6	5,368	1203	687	43%

structures [10], [18]. But eliminating redundancy of rules is not a focus of these approaches. The approaches proposed in [13] and [19] focus on extracting non-redundant rules. Both of them make use of the closure of the Galois connection [5] to extract non-redundant rules from frequent closed itemsets instead of from frequent itemsets. One difference between the two approaches is the definition of redundancy. The approach proposed in [19] extracts the rules with shorter antecedent and shorter consequent as well among rules which have the same confidence, while the method proposed in [13] defines that the non-redundant rules are those which have minimal antecedents and maximal consequents. Our definition to redundant rules is similar to that of [13]. However, the requirement to redundancy is relaxed, and the less requirement makes more rules to be considered redundant and thus eliminated. Most importantly, we prove that the elimination of such redundant rules does not reduce the belief to the extracted rules and the capacity of the extracted non-redundant rules for solving problems will also not be reduced. The concept of non-derivable itemsets was introduced in [4]. The basic idea is to find lower and upper bounds on the support of an itemset based on the support of its subsets. When these bounds are equal, the itemset is considered derivable. The set of frequent non-derivable itemsets allows for deriving the supports of all other frequent itemsets and as such forms a concise representation from which all other frequent itemsets can be derived. Goethals proposed a method to derive non-derivable rules from the non-derivable itemsets [6]. The amount of the non-derivable rules is much smaller than the size of the entire rule set. However, it was not discussed whether the non-derivable rule set has the same capacity to solve problems as the entire rule set.

VI. CONCLUSION

One challenge problem with association rule mining is the redundancy in the extracted rules. The work presented in this paper aims at improving the quality of association rules by eliminating redundancy. In this paper, we proposed a relaxed definition of redundancy and a concise representation of approximate association rules. We theoretically proved that the proposed Reliable Approximate basis can eliminate considerable amount of redundancy. Based on certainty factor theory, we also proved that the elimination of the redundancy using the proposed Reliable basis does not reduce the belief to the extracted rules. Similar to the Min-max basis, the proposed Reliable approximate basis is not only a concise but also a lossless representation of approximate rules. From the Reliable approximate basis, all approximate rules can be deduced.

REFERENCES

[1] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4:217–240, 2000.

[2] M. J. A. Berry and G. S. Linoff. *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley and Sons, 1997.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD Conference*, pages 255–264, 1997.

[4] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. pages 74–85. Springer, 2002.

[5] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1999.

[6] B. Goethals, J. Muhonen, and H. Toivonen. Mining non-derivable association rules. In *Proceedings of the SIAM International Conference on Data Mining*, pages 239–249, 2005.

[7] J. Han and Y. Fu. Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 11:798–804, 5 2000.

[8] J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-based multidimensional data mining. *IEEE Computer*, 32(8):46–50, 1999.

[9] M. Kryszkiewicz, H. Rybinski, and M. Gajek. Dataless transitions between concise representations of frequent patterns. *Journal of Intelligent Information Systems*, 22(1):41–70, 2004.

[10] Y. Li, W. Yang, and Y. Xu. Multi-tier granule mining for representations of multidimensional association rules. In *the 6th IEEE International Conference on Data Mining (ICDM06)*, pages 953–958, 2006.

[11] R. T. Ng, V. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning otimizations of constrained association rules. In *Proceedings of the SIGMOD conference*, pages 13–24, 1998.

[12] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rultes using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[13] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal. Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1):29–60, 2005.

[14] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3/4):351–379, 1975.

[15] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proceedings of the KDD Conference*, pages 67–73, 1997.

[16] R. Wille. *Restructuring lattices theory: An approach based on hierarchies of concepts*. I. Rival (editor), Ordered sets. Dordrecht-Boston, 1982.

[17] Y. Xu and Y. Li. Generating concise association rules. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM07)*, pages 781–790, 2007.

[18] W. Yang, Y. Li, J. Wu, and Y. Xu. Granule mining oriented data warehousing model for representations of multidimensional association rules. *International Journal of Intelligent Information and Database Systems*, 2008.

[19] M. J. Zaki. Generating non-redundent association rules. In *Proceedings of the KDD Conference*, pages 34–43, 2000.

[20] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248, 2004.